

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 FREE-MAD: CONSENSUS-FREE MULTI-AGENT DE- BATE

Anonymous authors

Paper under double-blind review

ABSTRACT

Multi-agent debate (MAD) is an emerging approach to improving the reasoning capabilities of large language models (LLMs). Existing MAD methods rely on multiple rounds of interaction among agents to reach consensus, and the final output is decided by majority voting in the last round. However, this consensus-based design faces several limitations. First, multiple rounds of communication increases token overhead and limits scalability. Second, due to the inherent conformity of LLMs, agents that initially produce correct responses may be influenced by incorrect ones during the debate process, causing error propagation. Third, majority voting introduces randomness and unfairness in the decision-making phase, and can degrade the reasoning performance. To address these issues, we propose FREE-MAD, an alternative and novel MAD framework that eliminates the need for consensus among agents. FREE-MAD introduces a novel score-based decision mechanism that evaluates the entire debate trajectory rather than relying on the last round only. This mechanism tracks how each agent’s reasoning evolves, enabling more accurate and fair outcomes. In addition, FREE-MAD reconstructs the debate phase by introducing anti-conformity, a mechanism that enables agents to mitigate excessive influence from the majority. Experiments on eight benchmark datasets demonstrate that FREE-MAD significantly improves reasoning performance while requiring only a single-round debate and thus reducing token costs. We also show that compared to existing MAD approaches, FREE-MAD exhibits improved robustness in real-world attack scenarios.

1 INTRODUCTION

Large language models (LLMs), due to their strong reasoning capabilities, have been widely applied in domains such as chatbots (Li et al., 2024a), programming (Yan et al., 2024), healthcare (Liu et al., 2024a), and cybersecurity (Zou et al., 2024). Recent applications of LLM agents (Luo et al., 2025b) have placed increasing demands on their reasoning performance. To improve the reasoning accuracy of LLM agents, many studies have explored training-free methods such as Chain-of-Thought (CoT) (Wei et al., 2022), self-refinement (Madaan et al., 2023) and self-consistency (Wang et al., 2023). However, these methods focus on the usage of single LLMs and lack collaboration among multiple models. In contrast, multi-agent debate (MAD) (Zeng et al., 2025; Du et al., 2024; Chan et al., 2024; Liu et al., 2025) has emerged as a prominent solution, showing that multiple agents engaged in multi-round interactions can achieve substantially better reasoning performance than a single agent. Indeed, MAD has been used in various scenarios such as software issue resolution (Li et al., 2025), mathematical reasoning (Zhang & Xiong, 2025), and code summarization (Chun et al., 2025).

Existing efforts to optimize MAD focus primarily on the reasoning strategies of individual agents during debates (Liu et al., 2025) or improving scalability (Zeng et al., 2025). Meanwhile, recent studies reveal that LLM agents can exhibit conformity (Weng et al., 2025; Zhu et al., 2025; Cho et al., 2025), meaning that during multi-agent interactions, agents tend to favor answers endorsed by the majority. In existing MAD frameworks, conformity-driven consensus (Sun et al., 2024; Zeng et al., 2025; Li et al., 2024b) is employed during the debate process to obtain the correct answer as the final decision (Chan et al., 2024). However, such consensus reduces reasoning accuracy. The consensus-based MAD schemes suffer from the Silent Agreement problem (Wang et al., 2025b). Even when the agents start with divergent opinions, they remain silent during the discussion due to conformity.

054 As a result, the agent group eventually provides an incorrect answer. More importantly, consensus
 055 naturally demands more debate rounds, increasing token consumption and limiting scalability.
 056

057 To address these limitations, we propose FREE-MAD, a consensus-free MAD framework that
 058 reconstructs both the debate stage and the decision stage. In the debate stage, we integrate the
 059 conventional conformity mode as used in prior work with a new mode called anti-conformity. In
 060 particular, the anti-conformity mode leverages CoT to encourage agents to identify flaws in the
 061 outputs from other agents; this situation is in contrast to existing approaches that use consensus
 062 as an indicator of correctness. In the decision stage, we propose a score-based mechanism that
 063 evaluates *all* intermediate outputs across debate rounds, instead of focusing solely on the final round
 064 as in traditional MAD frameworks. By tracking changes in the reasoning trajectories of all agents,
 065 FREE-MAD assigns scores to all candidate responses without requiring consensus in the debate
 066 stage. Furthermore, we theoretically formalize the MAD protocol, enabling a formal comparison
 067 between FREE-MAD and existing MAD variants.

068 To evaluate FREE-MAD, we have conducted extensive experiments on eight benchmark datasets,
 069 covering knowledge-based reasoning, logical reasoning, and mathematical reasoning with varying
 070 levels of difficulty. We show that FREE-MAD outpaces baseline approaches in terms of reasoning
 071 accuracy. In particular, FREE-MAD achieves improved accuracy with fewer debate rounds, thus
 072 accelerating the debate process while maintaining strong scalability. Furthermore, we show that
 073 FREE-MAD exhibits enhanced robustness (Chen et al., 2024a) and maintains accurate reasoning
 074 against communication attacks (He et al., 2025), where interactions among agents might be partially
 075 disrupted. We summarize our contributions as follows:

- 076 • We propose a novel consensus-free MAD framework, called FREE-MAD, with dedicated opti-
 077 mizations in both the debate and decision stages. We design a score-based decision mechanism
 078 that evaluates all intermediate results across debate rounds, enabling accurate reasoning without
 079 requiring consensus.
- 080 • We provide a syntax for the MAD protocol and perform a formal analysis of FREE-MAD and
 081 existing mechanisms.
- 082 • We implement FREE-MAD and conduct extensive experiments on eight benchmarks. We show
 083 that FREE-MAD outperforms existing protocols in terms of reasoning accuracy, scalability, and
 084 robustness.

085 2 RELATED WORK

086 **MAD Protocols.** When MAD was introduced, it was intended to improve reasoning by encouraging
 087 LLMs to reach consensus across their answers. However, such a consensus cannot be guaranteed in
 088 theory and is generally achievable only in practice (Du et al., 2024). Moreover, consensus becomes
 089 easier to achieve when agents adjust their trust between self-generated and externally generated
 090 content, yet this adjustment tends to reduce reasoning accuracy. The researchers did not investigate
 091 the deeper conformity issue that underlies this performance drop. Subsequent studies have focused
 092 on optimizing the debate stage of MAD (Chan et al., 2024; Chen et al., 2024b; Zeng et al., 2025;
 093 Liang et al., 2024; Liu et al., 2025). In these works, consensus is commonly treated as the default
 094 goal of the debate stage (Li et al., 2024b). In this paper, we show that MAD can operate effectively
 095 without requiring consensus.

096 **MAD Protocol Security.** In the debate stage, traditional consensus-oriented approaches for agents
 097 tend to lack robustness in unreliable network environments. Communication attacks (He et al., 2025)
 098 can force some agents to withdraw from the debate, preventing them from receiving others' responses.
 099 This delays consensus and increases overhead in adaptive schemes (Liang et al., 2024), ultimately
 100 reducing reasoning accuracy. Another line of work allows agents to return both their individual
 101 responses and a self-assessed confidence score (Chen et al., 2024b), which is then used in a weighted
 102 aggregation of the final result. However, due to LLM hallucinations (Ji et al., 2023), such confidence
 103 may be unreliable. Existing decision mechanisms in MAD are inadequate for addressing the security
 104 risks associated with LLM conformity. In real-world deployments, if a small subset of agents is
 105 compromised via prompt injection attacks (Greshake et al., 2023; Liu et al., 2024b; Zhan et al., 2025),
 106 the system may converge toward a shared but incorrect answer. When decisions are made using
 107 mechanisms such as majority voting, this can lead to a complete failure of the MAD system. Other

108 approaches (Liang et al., 2024) use an LLM-as-a-Judge (Zheng et al., 2023) framework, where an
 109 LLM decides the final outcome. This approach can produce biased results and is prone to conformity,
 110 making it effectively equivalent to majority voting. In addition, if the LLM is compromised by a
 111 prompt injection attack, the accuracy of the MAD system’s final output may drop significantly. We
 112 discuss related work on agent diversity in MAD further in Appendix B.

114 3 PRELIMINARY ANALYSIS

116 3.1 MAD PROTOCOL: A FORMAL TWO-PHASE DECOMPOSITION

118 To enable a formal analysis, we decompose the MAD protocol into two core stages: **Debate** and
 119 **Decision**. The Debate stage internally unfolds over R iterative rounds, culminating in a set of final
 120 answers. Formally, given a set of N agents denoted as $\{a_i\}_{i=1}^N$, the protocol is defined as:

$$122 \quad \{r_i^R\}_{i=1}^N \leftarrow \text{Debate}(\{a_i\}_{i=1}^N, q, p, R), \quad (1)$$

$$124 \quad r_{\text{final}} \leftarrow \text{Decide}(\{r_i^R\}_{i=1}^N). \quad (2)$$

125 In the **Debate** stage, all agents engage in a multi-round interaction based on a user query q and a
 126 guiding prompt p that specifies how agents should debate. The debate unfolds over R rounds. The
 127 debate begins with an initial step where each agent generates a preliminary response to q , which is
 128 then broadcast to all other agents as an auxiliary context (Yang et al., 2025). Subsequently, each
 129 agent a_i iteratively updates its own answer r_i^k over R rounds, resulting in a final set of responses
 130 $\{r_i^R\}_{i=1}^N$. The complete history of utterances up to round $k - 1$ is denoted as the context $C^{(k-1)}$. We
 131 model LLM agents as a probabilistic process to capture their generative behavior. In round k , agent
 132 a_i produces its response r_i^k by sampling from a conditional probability distribution defined over the
 133 preceding context and p :

$$134 \quad r_i^k \sim P_{a_i}(r|C^{(k-1)}, p). \quad (3)$$

136 To study the interplay between independent reasoning and conformity, we model the overall probabil-
 137 ity distribution P_{a_i} with a formulation that separates the contributions of the two factors:

$$139 \quad P_{a_i}(r|C^{(k-1)}, p) = \frac{1}{Z} \cdot P_{\text{in}}(r|q, p) \cdot \exp(\beta(p) \cdot S_{\text{con}}(r, C^{(k-1)})), \quad (4)$$

141 where the independent reasoning distribution $P_{\text{in}}(r | q, p)$ mathematically characterizes the agent’s
 142 intrinsic reasoning ability given the question q and prompt p . This ability excludes peer influence.
 143 The conformity score $S_{\text{con}}(r, C^{(k-1)})$ measures how much a candidate response r aligns with peer
 144 utterances in $C^{(k-1)}$. Its effect is scaled by the conformity parameter $\beta(p)$, which is determined
 145 by p . By default, LLMs exhibit a tendency to conform. We capture this by assuming $\beta_{\text{default}} > 0$
 146 when p is empty or neutral. If p encourages critical reasoning, $\beta(p)$ can be negative, acting as a
 147 regularizer that reduces alignment and pushes responses toward P_{in} . Conversely, if p encourages
 148 agreement with majority opinions, $\beta(p)$ assumes a larger positive value, which speeds up convergence
 149 towards consensus. Finally, the model combines all components via an exponential transformation
 150 and normalizes by Z to ensure a valid probability distribution.

151 In the **Decision** stage, a final output r_{final} is selected from $\{r_i^R\}_{i=1}^N$, typically through mechanisms
 152 such as majority voting. In this work, R excludes initial response generation and begins once agents
 153 start receiving responses from others.

155 3.2 WEAKNESSES OF EXISTING MAD APPROACHES

157 **Reasoning Accuracy.** MAD approaches (Du et al., 2024; Yang et al., 2025; Li et al., 2024b) design
 158 the decision stage¹ to operate on the final round’s N responses in the debate, while overlooking
 159 the remaining $R \times N$ intermediate responses that emerged throughout the debate process. This
 160 omission diminishes the influence of these earlier responses on determining r_{final} , thereby reducing

151 In this work, we focus exclusively on scenarios where only one option can be selected.

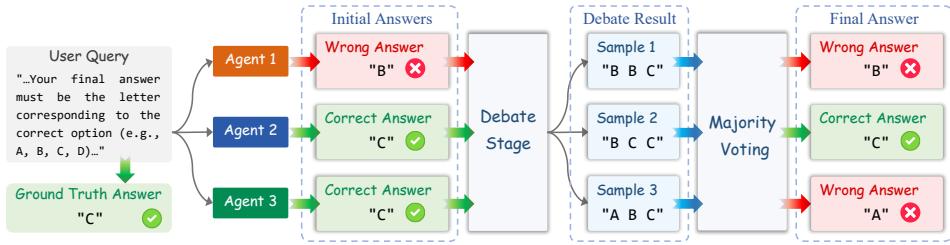


Figure 1: Existing MAD approaches may obtain final answers that are even less accurate than the initial ones.

both the accuracy and fairness of the final outcome. In consensus-based debates, the process ends once the agents reach agreement, even if the answer is incorrect (Chen et al., 2024b; Wang et al., 2025b). However, Du et al. (2024) has shown that it is still possible for the correct answer to emerge during later stages of the debate, even if none of the agents initially generates a correct answer. Early termination thus reduces MAD’s problem-solving accuracy. From empirical observations, we find that the initial responses generated independently by multiple agents may outperform the debate results obtained after applying MAD. As shown in Figure 1, applying majority voting directly to initial answers can yield the correct result, while debate outcomes may be incorrect. We illustrate three possibilities. In Sample 1, the agents reach consensus on an incorrect answer. Sample 2 represents the desired outcome, where the correct answer "C" holds the majority. Outcomes such as Sample 3 have received little attention. The set $\{r_i\}_{i=1}^N$ contains entirely distinct outputs, with no repetitions or equal counts for multiple answers. Under these circumstances, the final answer is determined either by selecting a response at random from the set $\{r_i\}_{i=1}^N$ or by choosing the first one. Both strategies substantially degrade the accuracy of MAD. Therefore, majority voting is unsuitable for decisions based on debate outcomes. More robust and practical mechanisms are required.

Robustness. Most existing multi-agent systems lack robustness (Chen et al., 2024a; Zhang et al., 2024; He et al., 2025). MAD is even more vulnerable to attacks (Qi et al., 2025) due to its consensus mechanism. Researchers (Luo et al., 2025a) mainly enhance the robustness of multi-agent systems by introducing blockchain and leveraging its traditional distributed consensus protocols. However, the use of blockchain, including smart contracts (Li et al., 2023) and consensus execution (Zhang et al., 2023; Duan et al., 2018), drastically impacts system performance.

4 METHODOLOGY

In this section, we first provide a formalization to introduce our proposed FREE-MAD (see Section 4.1). We then present the two core techniques within FREE-MAD: a debate protocol that operates without requiring consensus (see Section 4.2), and a score-based decision mechanism (see Section 4.3). Then, we provide a description of FREE-MAD (see Section 4.4). Finally, we conduct an analysis between FREE-MAD and existing approaches (see Section 4.5).

4.1 FREE-MAD

FREE-MAD focuses on the complete set of outcomes generated throughout the entire MAD process, rather than limiting attention to only the final-round responses, as done in traditional approaches. Our proposed debate protocol incorporates all responses into the decision-making process. This perspective can be formally represented using the following matrix formulation:

$$\text{Decide}[\alpha_0, \alpha_1, \dots, \alpha_{R-1}, \alpha_R] = \begin{bmatrix} r_1^0 & r_1^1 & \dots & r_1^{R-1} & r_1^R \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r_N^0 & r_N^1 & \dots & r_N^{R-1} & r_N^R \end{bmatrix} \rightarrow r_{\text{final}}. \quad (5)$$

The highlighted entries ($\alpha_0, \alpha_1, \dots, \alpha_{R-1}$) represent components that were not considered in previous MAD approaches. r_i^0 denotes the initial response generated by agent i . Unlike prior MAD frameworks that decouple debate and decision stages, FREE-MAD innovatively integrates decision-making into the debate phase, resulting in a unified process (see Figure 2).

216 4.2 CONSENSUS-FREE DEBATE
217

218 The agents in MAD exchange responses and update their answers based on a prefix prompt. A
219 common prompt is "The responses from other agents are as follows" (Du et al., 2024). Due to the
220 conformity of LLMs, such prompts often bias the agent toward the majority answer. This undermines
221 rigorous reasoning and raises the likelihood that reasoning errors remain unnoticed. To mitigate this
222 conformity, we design an additional CoT-based prompt (see Appendix E) appended after the prefix
223 prompt, motivated by (Weng et al., 2025). We provide agents with a scenario that includes adversarial
224 agents. We instruct agents to carefully assess the discrepancies between their own answers and those
225 from peers. Agents are expected to change their beliefs only if there is a clear indication that their
226 own answer is incorrect, rather than aiming to reach consensus with others. This mechanism is
227 intended to reduce the propagation of incorrect answers during the MAD process.
228

229 Our structured and critical reasoning prompt forms the core mechanism for optimizing the probabilis-
230 tic model (Equation 4) and has two main effects. First, it improves the quality of the independent
231 reasoning distribution P_{in} . FREE-MAD requires each agent to provide an answer along with a
232 detailed reasoning trace, which is incorporated into the next-round context $C^{(k-1)}$. Agent a_i uses
233 this context to analyze peers' reasoning rather than just their answers. If the reasoning behind the
234 majority answer is flawed, a_i 's own critical thinking will assign a very low probability to that answer
235 in P_{in} . Second, it balances conformity. Even if a popular but incorrect answer has a high S_{con} , a low
236 P_{in} keeps its overall probability P_{a_i} low.
237

Algorithm 1: MAD Protocol via Score-Based Decision and its Evaluation

238 **Input:** Answer matrix $A \in \mathbb{R}^{N \times (R+1)}$ from N agents over R rounds; List of task inputs and
239 ground truth responses $\mathcal{D}_{task} = \{q_i, r_i^g\}$; Weights $\mathcal{W} = \{w_i\}_{i=1}^4$; Guiding prompt p

240 **Output:** Accuracy

241 success $\leftarrow 0$; Initialize empty score dictionary $S \leftarrow \emptyset$

242 **for** q, r^g in \mathcal{D}_{task} **do**

243 **for** $k \leftarrow 0$ **to** R **do**

244 $f = (k + 1)^{-1}$ # Initial scoring factor with a non-zero value.

245 **for** $i \leftarrow 1$ **to** N **do**

246 **if** $k \neq 0$ **then**

247 Context $C \leftarrow$ Aggregate responses from other agents in round $k - 1$

248 $r_i^k \leftarrow P_{a_i}(q, p)$; Update C and A with r_i^k

249 $\hat{r} \leftarrow A[i][k]$

250 **if** $k = 0$ **then**

251 $S[\hat{r}] \leftarrow S[\hat{r}] + w_1 f$ # Assign an initial score to the answer.

252 **else**

253 $r_p \leftarrow A[i][k - 1]$ # Find the answer of agent a_i in the previous round.

254 **if** $\hat{r} \neq r_p$ **then**

255 **if** $r_p \in S$ **then**

256 $S[r_p] \leftarrow S[r_p] - w_2 f$ # The answer has been transferred from r_p to \hat{r} .

257 $S[\hat{r}] \leftarrow S[\hat{r}] + w_3 f$

258 **else**

259 $S[\hat{r}] \leftarrow S[\hat{r}] + w_4 f$ # The agent a_i maintains this answer.

260 Remove keys from S where key is None.

261 $MS \leftarrow \arg \max_k S[k]$

262 **if** $|MS| > 1$ **then**

263 Randomly choose $r_{final} \in MS$ # Guarantee robustness.

264 **else**

265 $r_{final} \leftarrow MS[0]$

266 success \leftarrow success + $\mathbb{I}(r_{final} = r^g)$

267 **Return** $\frac{\text{success}}{|\mathcal{D}|}$

268 4.3 SCORE-BASED DECISION MECHANISM

269 Based on the existing MAD framework (Du et al., 2024) and formalization of the traditional MAD
270 protocol (Subramaniam et al., 2025), we describe our protocol as shown in Algorithm 1. This method

270 maintains a matrix $A \in \mathbb{R}^{N \times (R+1)}$ that records the real-time responses of each agent across debate
 271 rounds. Concurrently, a score dictionary S is used to track the scores associated with multiple
 272 answers that emerge throughout the debate. The mechanism evaluates the likelihood of an answer
 273 being correct by tracking whether agents exhibit a shift in their opinions across rounds (whether the
 274 answer provided in the current round differs from that in the previous round). The agents in this
 275 framework are not designed to seek consensus; instead, they rigorously assess the reasoning behind
 276 the answers. Therefore, a change in an agent’s response is interpreted as an indicator that a more
 277 accurate answer has been identified. Specifically, answers that agents abandon are considered more
 278 likely to be incorrect, whereas newly adopted answers are treated as more likely to be correct. This
 279 dynamic informs the score updates within the dictionary S .

280 We assign different weights $w_i f$ to
 281 answers based on whether agents
 282 have changed their responses between
 283 rounds. Here, $w_i \in \mathcal{W}$ represents
 284 system-defined parameters, and f is
 285 a correction factor inversely propor-
 286 tional to the current round number.
 287 As the number of rounds increases,
 288 the amount of contextual infor-
 289 mation each agent receives also grows,
 thereby increasing the risk of confor-

290 mity, particularly for agents based on smaller models. To limit the influence of conformity, the
 291 impact of opinion shifts in later rounds is downweighted via the factor f . We define MS as the set
 292 of answers with the highest score. Although this set usually contains a single candidate, we adopt a
 293 randomized selection strategy to maintain theoretical robustness.

294

295

296 4.4 FRAMEWORK DESIGN

297

298 As noted in (Weng et al., 2025), LLMs’ con-
 299 formity presents a double-edged sword. On the one hand, it fosters consensus and cohesive out-
 300 comes. On the other hand, it undermines the reli-
 301 ability of agents’ judgments in sensitive domains
 302 such as voting. Debate based on anti-conformity
 303 mitigates the negative effects of blind confor-
 304 mity among agents. However, for relatively
 305 simple tasks, LLMs may show excessive anti-
 306 conformity, which causes stubbornness and re-
 307 duces reasoning accuracy. Therefore, we argue
 308 that conformity-oriented and anti-conformity-
 309 oriented debate modes should be adapted and
 310 switched according to the task. To achieve finer
 311 control over excessive conformity, the weight
 312 parameter \mathcal{W} in the score mechanism can be
 313 adjusted, which regulates the balance between
 314 an agent changing its opinion and maintaining
 315 its original stance. Based on this idea, we pro-
 316 pose FREE-MAD, which consists of two sub-schemes
 317 that share a score-based decision stage.
 318 FREE-MAD-N integrates an anti-conformity
 319 debate process, while FREE-MAD-C integrates a
 320 conformity-based debate process. Together, the
 321 schemes extend the framework’s applicability to a
 322 wider range of real-world scenarios.

323

324 4.5 ANALYSIS

325

326 We compare our proposed FREE-MAD framework with existing MAD approaches, as shown in
 327 Table 1. Below, we provide a detailed analysis of the advantages of FREE-MAD.

Table 1: A comprehensive comparison between our proposed approach and existing methods.

Approaches	Decision	Anti-Conformity	Fairness	Security
Sparse MADLi et al. (2024b)	Majority Voting	✗	✗	✗
Liang et al. (2024)	LLM-as-a-Judge	✗	✗	✗
ReConcile (Chen et al., 2024b)	Weighted Voting	✗	✓	✗
ChatEval (Chan et al., 2024)	Majority Voting	✗	✗	✗
DMAD (Liu et al., 2025)	Majority Voting	✗	✗	✓
SoM (Du et al., 2024)	Majority Voting	✗	✓	✓
FREE-MAD-N	score	✓	✓	✓
FREE-MAD-C	score	✗	✓	✓

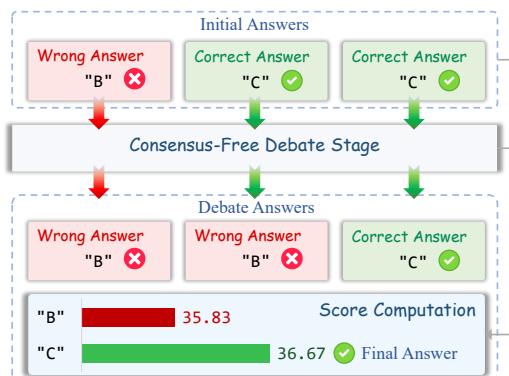


Figure 2: The inference process of our MAD framework. When the correct answers are in the minority in the final round, the framework is still able to identify the correct one as the final answer.

Figure 2 illustrates the inference process of our MAD framework. The process starts with 'Initial Answers' containing a 'Wrong Answer "B" (✗)' and two 'Correct Answer "C" (✓)'. These are processed through the 'Consensus-Free Debate Stage', which leads to 'Debate Answers' where both 'B' and 'C' are marked as wrong. Finally, 'Score Computation' is performed, showing scores: 'B' is 35.83 and 'C' is 36.67. The 'Final Answer' is 'C', demonstrating that the framework can identify the correct answer even when it is in the minority.

324 **Accuracy.** The FREE-MAD framework adopts a consensus-free debate strategy, which helps reduce
 325 the influence of conformity. This approach encourages agents to concentrate more on the correctness
 326 of reasoning steps, and may alleviate the effect of error propagation. In addition, existing methods
 327 usually require multiple rounds of debate to reach consensus. When the number of rounds R is small,
 328 such as when $R = 1$, these approaches may experience reduced reasoning performance due to the
 329 lack of consensus, which can lead to lower accuracy. In contrast, the performance of FREE-MAD is
 330 not closely tied to R .

331 **Scalability.** Prior work (Zeng et al., 2025) provides a general complexity analysis of the token cost
 332 (TC) in MAD as: $\mathcal{O}(NR^2V + N^2RV + NR|q|)$, where V is the maximum token cost for each agent.
 333 Therefore, a practical MAD framework should aim to achieve high accuracy with fewer agents and
 334 fewer rounds (Liu et al., 2025). The consensus-building process generally takes 2 to 3 rounds (Chen
 335 et al., 2024b; Du et al., 2024; Yang et al., 2025; Xiong et al., 2023) to be effective. In contrast, our
 336 method theoretically requires only a single round of debate without any consensus constraint, which
 337 substantially reduces token consumption.

338 **Security.** Under communication attacks, agents that withdraw from the debate process generally
 339 trigger only $S[\hat{r}] \leftarrow S[\hat{r}] + w_4f$ in Algorithm 1 under our score scheme, because these agents retain
 340 context containing only their own prior responses, which does not affect the overall debate process. In
 341 addition, the score strategy is executed entirely outside the LLM reasoning and follows a deterministic
 342 protocol, rendering it immune to LLM hallucination.

343 **Fairness.** During the debate stage, some approaches adopt role-based debate strategies (Chan et al.,
 344 2024), where agents are assigned unequal statuses and perform different functions. This design
 345 reinforces the implicit biases of LLMs (Vasista et al., 2025; Myung et al., 2025; Kim et al., 2024) and
 346 undermines the fairness of MAD systems (Xiong et al., 2023). In contrast, agents in FREE-MAD do
 347 not require any predefined roles, and all agents participate equally in the debate process.

348

349 5 EXPERIMENTS

350

351 5.1 EXPERIMENTAL SETUP

352

353 **Evaluation Benchmark.** Based on the comparison in Table 1 and the analysis in Section 4.5, we
 354 select the SoM framework (Du et al., 2024) as the baseline for our experiments to ensure a fair
 355 comparison (SoM is also widely adopted as a baseline in related work (Chen et al., 2024b; Wang
 356 et al., 2025a; Li et al., 2024b)). Moreover, our proposed FREE-MAD is implemented on top of SoM
 357 to minimize the influence of confounding factors. In Algorithm 1, the weights \mathcal{W} are initialized
 358 to $\{20, 25, 30, 20\}$ based on theoretical analysis. We present the experimental setup for evaluating
 359 the security of FREE-MAD in Appendix A. For the ablation study, we compare four schemes, as
 360 summarized in Table 2, including our FREE-MAD and the SoM baseline. This comparison highlights
 361 the effectiveness of the two core modules we developed.

362 **Datasets.** To comprehensively evaluate the capability of FREE-MAD, we
 363 conduct experiments on 8 benchmark
 364 datasets. For mathematical reasoning,
 365 we use GSM-Ranges (Shrestha et al.,
 366 2025) (levels 4 and 6), AIME2024,
 367 AIME2025 (Art of Problem Solving, 2025), and MATH500 (Lightman et al., 2024). For logical
 368 reasoning, we employ StrategyQA (Geva et al., 2021) and the Logical Fallacies dataset of MMLU
 369 (Hendrycks et al., 2021). For knowledge and theoretical reasoning, we adopt the multiple-choice ques-
 370 tions dataset from AICrypto (Wang et al., 2025c), which constitutes the first benchmark specifically
 371 constructed to assess the cryptographic capabilities of LLMs.

372

373 **Agent Groups.** To ensure that the MAD framework possesses the basic capability to handle our
 374 datasets, we design two configurations of MAD. For AIME2024 and AIME2025, we construct
 375 MAD with $N = 3$ based on Qwen1.5-7B-Chat² and DeepSeek-V3 (DeepSeek-AI et al., 2024).
 376 For the other datasets, we uniformly construct MAD with $N = 4$ using Qwen1.5-7B-Chat and
 377 Qwen2.5-72B-Instruct (Qwen et al., 2025). Other details are provided in Appendix.

Table 2: Module configurations of multiple comparative variants in ablation experiments.

Schemes	FREE-MAD-N	FREE-MAD-C	Baseline 1	Baseline 2 (SoM)
Debate Decision	Anti-conformity Score	Conformity Score	Anti-conformity Majority Voting	Conformity Majority Voting

²<https://qwenlm.github.io/zh/blog/qwen1.5>

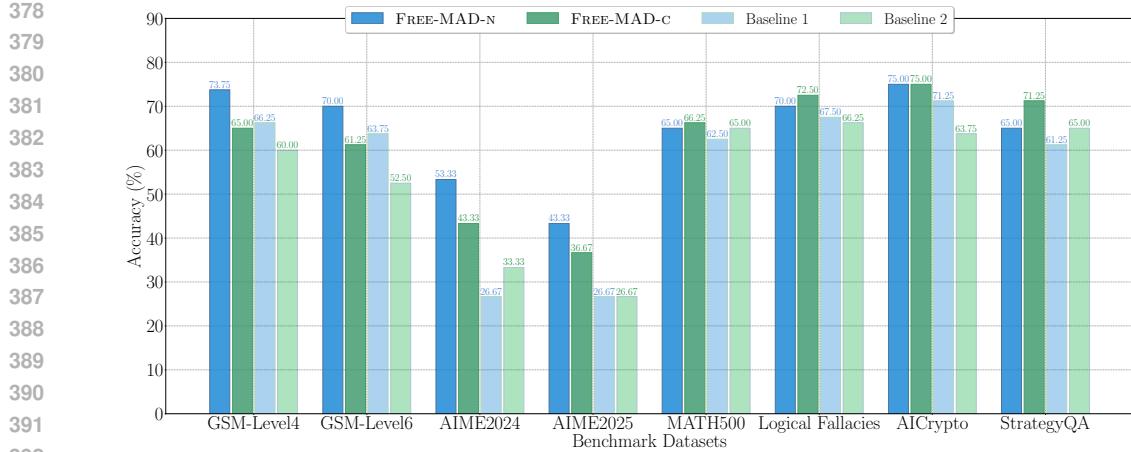


Figure 4: Comprehensive comparative experimental results for MAD frameworks across multiple benchmarks.

5.2 EVALUATION METRICS

To evaluate the reasoning performance and scalability of MAD, we assess both inference accuracy and token consumption. Following (Zeng et al., 2025), we adopt token consumption as the metric for scalability. The computation of accuracy follows Algorithm 1, while the calculation of token consumption is defined as follows:

$$TC = \sum_{k=0}^R \sum_{i=1}^N \mathcal{T}_k^i, \quad (6)$$

where \mathcal{T}_k^i denotes the number of output tokens generated by agent a_i in the k -th round.

6 MAIN RESULTS

6.1 REASONING PERFORMANCE

The evaluation results of reasoning accuracy on eight benchmarks are presented in Figure 4 and Table 4. Overall, our proposed FREE-MAD substantially outperforms the baselines, achieving average improvements of 13.0% and 16.5% over baselines, respectively. These results demonstrate a significant enhancement in reasoning performance. In particular, for mathematical reasoning tasks, the advantage of FREE-MAD becomes more evident with increasing problem difficulty. Reducing conformity clearly improves the effectiveness of MAD. The specific effects of conformity and anti-conformity on the reasoning process of LLMs are detailed in Appendix F. Notably, under anti-conformity, the reasoning process of LLMs appears to be more rational.

On the MATH500 dataset, we observe that weaker models exhibit a previously mentioned tendency toward rigidity in reasoning when conformity is suppressed. In such cases, these models fail to switch to correct lines of reasoning, resulting in comparable performance between FREE-MAD and the baselines. This limitation is expected, as a fixed set of agents cannot be universally optimal across all task categories. For logical and knowledge-based reasoning, FREE-MAD still significantly

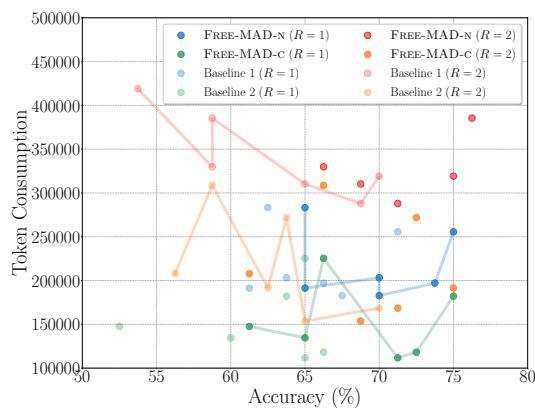


Figure 3: Comparison of token consumption and reasoning accuracy between our proposed schemes and baselines under different debate rounds.

432 Table 3: Comparison of reasoning accuracy between our schemes and baselines when $R = 1$.
433

Schemes	FREE-MAD-N				FREE-MAD-C				Baseline 1	Baseline 2
Accuracy	64.43% ($\uparrow 16\% / 19\%$)				61.41% ($\uparrow 10\% / 14\%$)				55.73%	54.06%
	20	40	60	80	100					
FREE-MAD-N	67.50	65.00	71.25	80.00		FREE-MAD-N	86.25	73.75	81.25	82.50
FREE-MAD-C	72.50	65.00	73.75	76.25		FREE-MAD-C	75.00	65.00	80.00	82.50
Baseline 1	47.50	42.50	48.75	66.25		Baseline 1	65.00	58.75	20.00	17.50
Baseline 2	41.25	48.75	60.00	52.50		Baseline 2	62.50	62.50	21.25	18.75
	GSM-Level4	MATH500	Logical Fallacies	AI-Crypto		GSM-Level4	MATH500	Logical Fallacies	AI-Crypto	

448 (a) Experimental results when $R = 1$. (b) Experimental results when $R = 2$.
449450 Figure 5: Empirical evaluation of the security of MAD frameworks across multiple benchmarks,
451 showing the comparison of their reasoning accuracy under communication attacks across varying
452 numbers of rounds.

453 outperforms the baselines. However, FREE-MAD-C consistently achieves better results than FREE-
454 MAD-N, because for models lacking relevant knowledge, idea switching under anti-conformity tends
455 to occur with relatively high randomness. Consequently, in such scenarios, conformity may lead to
456 more effective outcomes. By comparing the four variants in our ablation study, we demonstrate that
457 the proposed core score-based decision mechanism exhibits superior performance.

459

6.2 SCALABILITY

461 We compared the token consumption and reasoning accuracy of four MAD variants under $R = 1$
462 and $R = 2$, as shown in Figure 3. With an increasing number of debate rounds, the reasoning
463 accuracy of MAD improves. Notably, FREE-MAD achieves accuracy comparable to or even higher
464 than the two-round baseline 2 setting with only a single debate round, while incurring almost no
465 additional token consumption, demonstrating promising scalability. Specifically, FREE-MAD-N
466 achieves stronger reasoning accuracy compared to the baselines, whereas FREE-MAD-C offers better
467 scalability. More importantly, our approach eliminates the need for multiple debate rounds to reach
468 consensus, which significantly reduces the execution time of the MAD system.

470

6.3 SECURITY

472 Compared to the normal scenario, both baseline 1 and baseline 2 exhibit a substantial drop in accuracy,
473 reaching up to 20% (see Figure 5). In contrast, FREE-MAD consistently maintains very high accuracy.
474 Interestingly, in some cases, it even slightly outperforms the original accuracy. This behavior can
475 be attributed to the fact that communication attacks prevent some agents from receiving responses
476 from others, while simultaneously reducing the probability of receiving incorrect information. These
477 results demonstrate that FREE-MAD possesses strong robustness and security.

478

7 CONCLUSION

481 This paper proposes FREE-MAD, a novel MAD framework that integrates controllable conformity
482 with a score-based decision mechanism. Unlike traditional MAD approaches, FREE-MAD does
483 not rely on multi-round interactions or need to reach a consensus. By evaluating the entire debate
484 trajectory instead of relying solely on the final round, the accuracy of deciding the answer is enhanced.
485 Extensive experiments show that FREE-MAD outperforms existing MAD approaches in terms of
486 reasoning accuracy, scalability, and robustness.

486 REFERENCES
487

488 Art of Problem Solving. AIME Problems and Solutions, 2025. URL https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions. Accessed: 2025-05-15.

491 Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and
492 Zhiyuan Liu. Chateval: Towards better LLM-based evaluators through multi-agent debate. In
493 *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=FQepisCUwu>.

496 Bei Chen, Gaolei Li, Xi Lin, Zheng Wang, and Jianhua Li. Blockagents: Towards byzantine-robust
497 llm-based multi-agent coordination via blockchain. In *Proceedings of the ACM Turing Award
498 Celebration Conference - China 2024*, ACM-TURC '24, pp. 187–192, New York, NY, USA, 2024a.
499 Association for Computing Machinery. ISBN 9798400710117. doi: 10.1145/3674399.3674445.
500 URL <https://doi.org/10.1145/3674399.3674445>.

501 Justin Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-table conference im-
502 proves reasoning via consensus among diverse LLMs. In Lun-Wei Ku, Andre Martins, and
503 Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Com-
504 putational Linguistics (Volume 1: Long Papers)*, pp. 7066–7085, Bangkok, Thailand, August
505 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.381. URL
506 <https://aclanthology.org/2024.acl-long.381/>.

507 508 Young-Min Cho, Sharath Chandra Guntuku, and Lyle Ungar. Herd behavior: Investigating peer
509 influence in llm-based multi-agent systems, 2025. URL <https://arxiv.org/abs/2505.21588>.

511 512 Jina Chun, Qihong Chen, Jiawei Li, and Iftekhar Ahmed. Is multi-agent debate (mad) the silver
513 bullet? an empirical analysis of mad in code summarization and translation. *arXiv preprint
arXiv:2503.12029*, 2025.

514 515 DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang
516 Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli
517 Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen,
518 Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding,
519 Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi
520 Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song,
521 Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang,
522 Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan
523 Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang,
524 Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi
525 Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li,
526 Shanghao Lu, Shangyan Zhou, Shanhua Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye,
527 Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang,
528 Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanja Zhao, Wei An, Wen Liu,
529 Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyu Jin, Xianzu Wang,
530 Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha
531 Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu,
532 Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su,
533 Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong
534 Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng,
535 Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan
536 Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue
537 Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo,
538 Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu,
539 Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou,
Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu
Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan.
Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

540 Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving
 541 factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st*
 542 *International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

543

544 Sisi Duan, Michael K. Reiter, and Haibin Zhang. Beat: Asynchronous bft made practical. In
 545 *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*,
 546 CCS '18, pp. 2028–2041, New York, NY, USA, 2018. Association for Computing Machinery.
 547 ISBN 9781450356930. doi: 10.1145/3243734.3243812. URL <https://doi.org/10.1145/3243734.3243812>.

548

549 Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle
 550 use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions*
 551 *of the Association for Computational Linguistics*, 9:346–361, 2021. doi: 10.1162/tacl_a_00370.
 552 URL <https://aclanthology.org/2021.tacl-1.21/>.

553

554 Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario
 555 Fritz. Not what you've signed up for: Compromising real-world llm-integrated applications with
 556 indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and*
 557 *Security*, AISeC '23, pp. 79–90, New York, NY, USA, 2023. Association for Computing Machinery.
 558 ISBN 9798400702600. doi: 10.1145/3605764.3623985. URL <https://doi.org/10.1145/3605764.3623985>.

559

560 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,
 561 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms
 562 via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

563

564 Pengfei He, Yuping Lin, Shen Dong, Han Xu, Yue Xing, and Hui Liu. Red-teaming LLM multi-agent
 565 systems via communication attacks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and
 566 Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL*
 567 2025, pp. 6726–6747, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN
 568 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.349. URL <https://aclanthology.org/2025.findings-acl.349/>.

569

570 Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob
 571 Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International*
572 Conference on Learning Representations (ICLR), 2021.

573

574 Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating LLM
 575 hallucination via self reflection. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of*
 576 *the Association for Computational Linguistics: EMNLP 2023*, pp. 1827–1843, Singapore, December
 577 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.123.
 578 URL <https://aclanthology.org/2023.findings-emnlp.123/>.

579

580 Alex Kim, Keonwoo Kim, and Sangwon Yoon. DEBATE: Devil's advocate-based assessment and
 581 text evaluation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the*
 582 *Association for Computational Linguistics: ACL 2024*, pp. 1885–1897, Bangkok, Thailand, August
 583 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.112. URL
 584 <https://aclanthology.org/2024.findings-acl.112/>.

585

586 Han Li, Yuling Shi, Shaoxin Lin, Xiaodong Gu, Heng Lian, Xin Wang, Yantao Jia, Tao Huang, and
 587 Qianxiang Wang. Swe-debate: Competitive multi-agent debate for software issue resolution. *arXiv*
 588 *preprint arXiv:2507.23348*, 2025.

589

590 Huizhong Li, Yujie Chen, Xiang Shi, Xingqiang Bai, Nan Mo, Wenlin Li, Rui Guo, Zhang Wang, and
 591 Yi Sun. Fisco-bcos: An enterprise-grade permissioned blockchain system with high-performance.
 592 In *Proceedings of the International Conference for High Performance Computing, Networking,*
 593 *Storage and Analysis*, SC '23, New York, NY, USA, 2023. Association for Computing Machinery.
 594 ISBN 9798400701092. doi: 10.1145/3581784.3607053. URL <https://doi.org/10.1145/3581784.3607053>.

595

596 Weitao Li, Junkai Li, Weizhi Ma, and Yang Liu. Citation-enhanced generation for LLM-based
 597 chatbots. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd*

594 *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp.
 595 1451–1466, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.
 596 18653/v1/2024.acl-long.79. URL <https://aclanthology.org/2024.acl-long.79/>.
 597

598 Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie.
 599 Improving multi-agent debate with sparse communication topology. In Yaser Al-Onaizan,
 600 Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational
 601 Linguistics: EMNLP 2024*, pp. 7281–7294, Miami, Florida, USA, November 2024b. Asso-
 602 ciation for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.427. URL
 603 <https://aclanthology.org/2024.findings-emnlp.427/>.
 604

604 Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi,
 605 and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent
 606 debate. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024
 607 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904, Miami,
 608 Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/
 609 2024.emnlp-main.992. URL <https://aclanthology.org/2024.emnlp-main.992/>.
 610

610 Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan
 611 Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *The Twelfth
 612 International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=v8L0pN6EOi>.
 613

614 Lei Liu, Xiaoyan Yang, Junchi Lei, Yue Shen, Jian Wang, Peng Wei, Zhixuan Chu, Zhan Qin, and
 615 Kui Ren. A survey on medical large language models: Technology, application, trustworthiness,
 616 and future directions. *arXiv preprint arXiv:2406.03712*, 2024a.
 617

618 Yexiang Liu, Jie Cao, Zekun Li, Ran He, and Tieniu Tan. Breaking mental set to improve reasoning
 619 through diverse multi-agent debate. In *The Thirteenth International Conference on Learning
 620 Representations*, 2025. URL <https://openreview.net/forum?id=t6QHYUOQL7>.
 621

621 Yupei Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. Formalizing and bench-
 622 marking prompt injection attacks and defenses. In *33rd USENIX Security Symposium (USENIX
 623 Security 24)*, pp. 1831–1847, Philadelphia, PA, August 2024b. USENIX Association. ISBN 978-1-
 624 939133-44-1. URL <https://www.usenix.org/conference/usenixsecurity24/presentation/liu-yupe>.
 625

626 Haoxiang Luo, Gang Sun, Yinqi Liu, Dongcheng Zhao, Dusit Niyato, Hongfang Yu, and Schahram
 627 Dustdar. A weighted byzantine fault tolerance consensus driven trusted multiple large language
 628 models network. *arXiv preprint arXiv:2505.05103*, 2025a.
 629

630 Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi
 631 Chen, Ziyue Qiao, Qingqing Long, et al. Large language model agent: A survey on methodology,
 632 applications and challenges. *arXiv preprint arXiv:2503.21460*, 2025b.
 633

633 Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon,
 634 Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder,
 635 Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: iterative
 636 refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural
 637 Information Processing Systems*, NIPS ’23, Red Hook, NY, USA, 2023. Curran Associates Inc.
 638

638 Junho Myung, Yeon Su Park, Sunwoo Kim, Shin Yoo, and Alice Oh. PapersPlease: A bench-
 639 mark for evaluating motivational values of large language models based on ERG theory. In
 640 Ofir Ariv, Miruna Clinciu, Kaustubh Dhole, Rotem Dror, Sebastian Gehrmann, Eliya Habba,
 641 Itay Itzhak, Simon Mille, Yotam Perlitz, Enrico Santus, João Sedoc, Michal Shmueli Scheuer,
 642 Gabriel Stanovsky, and Oyvind Tafjord (eds.), *Proceedings of the Fourth Workshop on Gen-
 643 eration, Evaluation and Metrics (GEM²)*, pp. 522–531, Vienna, Austria and virtual meet-
 644 ing, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-261-9. URL
 645 <https://aclanthology.org/2025.gem-1.47/>.
 646

646 Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin
 647 Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity’s last exam. *arXiv preprint
 648 arXiv:2501.14249*, 2025.

648 Senmao Qi, Yifei Zou, Peng Li, Ziyi Lin, Xiuzhen Cheng, and Dongxiao Yu. Amplified vulnerabilities:
 649 Structured jailbreak attacks on llm-based multi-agent debate. *arXiv preprint arXiv:2504.16489*,
 650 2025.

651 Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan
 652 Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang,
 653 Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin
 654 Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi
 655 Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan,
 656 Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL
 657 <https://arxiv.org/abs/2412.15115>.

658 Safal Shrestha, Minwu Kim, and Keith Ross. Mathematical reasoning in large language mod-
 659 els: Assessing logical and arithmetic errors across wide numerical ranges. *arXiv preprint*
 660 *arXiv:2502.08680*, 2025.

661 Vighnesh Subramaniam, Yilun Du, Joshua B. Tenenbaum, Antonio Torralba, Shuang Li, and
 662 Igor Mordatch. Multiagent finetuning: Self improvement with diverse reasoning chains. In
 663 *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=JtGPIZpOrz>.

664 Qiushi Sun, Zhangyue Yin, Xiang Li, Zhiyong Wu, Xipeng Qiu, and Lingpeng Kong. Corex: Pushing
 665 the boundaries of complex reasoning through multi-model collaboration. In *ICLR 2024 Workshop*
 666 *on Large Language Model (LLM) Agents*, 2024. URL <https://openreview.net/forum?id=cCQRjp0ru5>.

667 Ishwara Vasista, Imran Mirza, Cole Huang, Rohan Rajasekhara Patil, Aslihan Akalin, Kevin Zhu,
 668 and Sean O'Brien. MALIBU benchmark: Multi-agent LLM implicit bias uncovered. In *ICLR*
 669 *2025 Workshop on Building Trust in Language Models and Applications*, 2025. URL <https://openreview.net/forum?id=iilhN2MycO>.

670 Haotian Wang, Xiyuan Du, Weijiang Yu, Qianglong Chen, Kun Zhu, Zheng Chu, Lian Yan, and
 671 Yi Guan. Learning to break: Knowledge-enhanced reasoning in multi-agent debate system.
 672 *Neurocomputing*, 618:129063, 2025a.

673 Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha
 674 Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language
 675 models. In *The Eleventh International Conference on Learning Representations*, 2023. URL
 676 <https://openreview.net/forum?id=1PL1NIMMrw>.

677 Yihan Wang, Qiao Yan, Zhenghao Xing, Lihao Liu, Junjun He, Chi-Wing Fu, Xiaowei Hu, and
 678 Pheng-Ann Heng. Silence is not consensus: Disrupting agreement bias in multi-agent llms via
 679 catfish agent for clinical decision making. *arXiv preprint arXiv:2505.21503*, 2025b.

680 Yu Wang, Yijian Liu, Liheng Ji, Han Luo, Wenjie Li, Xiaofei Zhou, Chiyun Feng, Puji Wang, Yuhan
 681 Cao, Geyuan Zhang, et al. Aicrypto: A comprehensive benchmark for evaluating cryptography
 682 capabilities of large language models. *arXiv preprint arXiv:2507.09580*, 2025c.

683 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V
 684 Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language mod-
 685 els. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Ad-
 686 vances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Asso-
 687 ciates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

688 Zhiyuan Weng, Guikun Chen, and Wenguan Wang. Do as we do, not as you think: the conformity of
 689 large language models. In *The Thirteenth International Conference on Learning Representations*,
 690 2025. URL <https://openreview.net/forum?id=st77ShxP1K>.

691 Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. Examining inter-consistency of large
 692 language models collaboration: An in-depth analysis via debate. In *The 2023 Conference on*
 693 *Empirical Methods in Natural Language Processing*, 2023. URL <https://openreview.net/forum?id=XEwQ1fDbDN>.

702 Weixiang Yan, Haitian Liu, Yunkun Wang, Yunzhe Li, Qian Chen, Wen Wang, Tingyu Lin, Weishan
 703 Zhao, Li Zhu, Hari Sundaram, and Shuguang Deng. CodeScope: An execution-based multilingual
 704 multitask multidimensional benchmark for evaluating LLMs on code understanding and generation.
 705 In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual*
 706 *Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5511–
 707 5558, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/
 708 v1/2024.acl-long.301. URL <https://aclanthology.org/2024.acl-long.301/>.

709 Yongjin Yang, Euiin Yi, Jongwoo Ko, Kimin Lee, Zhijing Jin, and Se-Young Yun. Revisiting
 710 multi-agent debate as test-time scaling: A systematic study of conditional effectiveness. *arXiv*
 711 *preprint arXiv:2505.22960*, 2025.

712 Yuting Zeng, Weizhe Huang, Lei Jiang, Tongxuan Liu, XiTai Jin, Chen Tianying Tiana, Jing Li, and
 713 Xiaohua Xu. S²-MAD: Breaking the token barrier to enhance multi-agent debate efficiency. In
 714 Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations*
 715 *of the Americas Chapter of the Association for Computational Linguistics: Human Language*
 716 *Technologies (Volume 1: Long Papers)*, pp. 9393–9408, Albuquerque, New Mexico, April 2025.
 717 Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.
 718 naacl-long.475. URL <https://aclanthology.org/2025.naacl-long.475/>.

719 Qiusi Zhan, Richard Fang, Henil Shalin Panchal, and Daniel Kang. Adaptive attacks break de-
 720 fenses against indirect prompt injection attacks on LLM agents. In Luis Chiruzzo, Alan Rit-
 721 ter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL*
 722 2025, pp. 7101–7117, Albuquerque, New Mexico, April 2025. Association for Computational
 723 Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.395. URL
 724 <https://aclanthology.org/2025.findings-naacl.395/>.

725 Boyang Zhang, Yicong Tan, Yun Shen, Ahmed Salem, Michael Backes, Savvas Zannettou, and Yang
 726 Zhang. Breaking agents: Compromising autonomous llm agents through malfunction amplification.
 727 *arXiv preprint arXiv:2407.20859*, 2024.

728 Haibin Zhang, Sisi Duan, Boxin Zhao, and Liehuang Zhu. Waterbear: practical asynchronous bft
 729 matching security guarantees of partially synchronous bft. In *Proceedings of the 32nd USENIX*
 730 *Conference on Security Symposium, SEC '23, USA, 2023*. USENIX Association. ISBN 978-1-
 731 939133-37-3.

732 Shaowei Zhang and Deyi Xiong. Debate4MATH: Multi-agent debate for fine-grained reasoning
 733 in math. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar
 734 (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16810–16824,
 735 Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-
 736 5. doi: 10.18653/v1/2025.findings-acl.862. URL <https://aclanthology.org/2025.findings-acl.862/>.

737 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,
 738 Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion
 739 Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann,
 740 A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Infor-
 741 mation Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023.
 742 URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

743 Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Collier, and Andreas Vlachos. Conformity in large
 744 language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher
 745 Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational*
 746 *Linguistics (Volume 1: Long Papers)*, pp. 3854–3872, Vienna, Austria, July 2025. Association for
 747 Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.195. URL
 748 <https://aclanthology.org/2025.acl-long.195/>.

749 Yuwen Zou, Yang Hong, Jingyi Xu, Lekun Liu, and Wenjun Fan. Leveraging large language models
 750 for challenge solving in capture-the-flag. In *2024 IEEE 23rd International Conference on Trust,
 751 Security and Privacy in Computing and Communications (TrustCom)*, pp. 1541–1550, 2024. doi:
 752 10.1109/TrustCom63139.2024.00213.

756 A FREE-MAD UNDER COMMUNICATION ATTACKS AND ITS EVALUATION
757

758 For the security evaluation, we construct MAD under communication attacks based on Algorithm 2,
759 and apply the same modification to the SoM framework to serve as a baseline for comparison.
760 Specifically, in the modified setting, we perform the operation of aggregating responses from other
761 agents only for the Context C of agents that are not under attack. The compromised agent is unable
762 to receive responses from other agents, while the other agents can still receive the outputs generated
763 by this agent. We evaluate the accuracy of MAD across multiple benchmark datasets by setting the
764 proportion of compromised agents to $|\mathcal{V}|/|\mathcal{N}| = 50\%$ (see Algorithm 2), which better reflects the
765 adversarial capability in real-world deployments. We evaluate the security of the MAD framework
766 under communication attacks on four datasets: GSM-Ranges (Level 4), MATH500, Logical Fallacies,
767 and AICrypto.

768 **Algorithm 2:** FREE-MAD under Communication Attacks
769

770 **Input:** Answer matrix $A \in \mathbb{R}^{N \times (R+1)}$ from N agents over R rounds; List of task inputs and
771 ground truth responses $\mathcal{D}_{\text{task}} = \{q_i, r_i^g\}$; Weights $\mathcal{W} = \{w_i\}_{i=1}^4$; Guiding prompt p ;
772 Agents under attack $\mathcal{V} = \{v_i\}_{i=1}^L$

773 **Output:** Accuracy
774 $\text{success} \leftarrow 0$; Initialize empty score dictionary $S \leftarrow \emptyset$
775 **for** q, r^g in $\mathcal{D}_{\text{task}}$ **do**
776 **for** $k \leftarrow 0$ **to** R **do**
777 $f = (k + 1)^{-1}$ *# Initial scoring factor with a non-zero value.*
778 **for** $i \leftarrow 1$ **to** N **do**
779 **if** $k \neq 0$ and $a_i \notin \mathcal{V}$ **then**
780 | Context $C \leftarrow$ Aggregate responses from other agents in round $k - 1$
781 | $r_i^k \leftarrow P_{a_i}(q, p)$; Update C and A with r_i^k
782 | $\hat{r} \leftarrow A[i][k]$
783 | **if** $k = 0$ **then**
784 | | $S[\hat{r}] \leftarrow S[\hat{r}] + w_1 f$ *# Assign an initial score to the answer.*
785 | **else**
786 | | $r_p \leftarrow A[i][k - 1]$ *# Find the answer of agent a_i in the previous round.*
787 | | **if** $\hat{r} \neq r_p$ **then**
788 | | **if** $r_p \in S$ **then**
789 | | $S[r_p] \leftarrow S[r_p] - w_2 f$ *# The answer has been transferred from r_p to \hat{r} .*
790 | | $S[\hat{r}] \leftarrow S[\hat{r}] + w_3 f$
791 | | **else**
792 | | $S[\hat{r}] \leftarrow S[\hat{r}] + w_4 f$ *# The agent a_i maintains this answer.*
793 | Remove keys from S where key is None.
794 | $MS \leftarrow \arg \max_k S[k]$
795 | **if** $|MS| > 1$ **then**
796 | | Randomly choose $r_{\text{final}} \in MS$ *# Guarantee robustness.*
797 | **else**
798 | | $r_{\text{final}} \leftarrow MS[0]$
799 | $\text{success} \leftarrow \text{success} + \mathbb{I}(r_{\text{final}} = r^g)$
800 | $\text{Accuracy} \leftarrow \frac{\text{success}}{|\mathcal{D}|}$
801 | **Return** Accuracy

802 B ADDITIONAL RELATED WORK
803

804 **Agent Diversity in MAD.** In previous studies, although some MAD frameworks have considered
805 heterogeneity and diversity among agents, their experimental evaluations were typically conducted
806 using LLMs with similar model sizes (Yang et al., 2025), such as Llama3.1-8B³, Qwen2.5-7B⁴, and
807 Gemma-2-9B⁵. This setup significantly limits the effectiveness of MAD and, in certain cases, results
808 in worse performance than self-consistency approaches. In real-world deployments, however, we aim

809 ³<https://huggingface.co/meta-llama/Llama-3.1-8B>

⁴<https://huggingface.co/Qwen/Qwen2.5-7B>

⁵<https://huggingface.co/google/gemma-2-9b>

810 to enable collaboration among diverse LLMs, each possessing different strengths (Chan et al., 2024;
 811 Liu et al., 2025), to act as equal peers within a MAD framework. Such collaboration is expected
 812 to outperform the strongest single agent on the same task and to solve a broader range of problems
 813 through cooperation, analogous to human team-based collaboration.

814 However, when the participating agents in a heterogeneous MAD system exhibit substantial differences
 815 in their capabilities, it may lead to significant variance in their confidence regarding their own
 816 responses. As a consequence of conformity in LLMs (Weng et al., 2025), agents may tend to adopt
 817 the outputs of peers, even when those outputs are incorrect. While conformity can be beneficial in
 818 facilitating consensus, it also introduces detrimental effects that require external mitigation.

819 **Consensus in MAD.** Reaching consensus was the fundamental objective when MAD was first
 820 proposed (Du et al., 2024). It is also a necessary condition for obtaining a correct final answer. To
 821 the best of our knowledge, all existing MAD methods adopt consensus as a core mechanism in their
 822 underlying design (Li et al., 2024b). Chan et al. (2024) does not explicitly require agents to reach
 823 consensus during the debate stage. However, it still applies majority voting in the decision phase,
 824 which preserves the essential logic of consensus.

825 C EXPERIMENTAL SETUP

826
 827
 828 In our experiments, we selected eight datasets to comprehensively cover different types of tasks.
 829 Regarding the number of samples, we referred to the configuration used in prior work (Liu et al., 2025;
 830 Du et al., 2024). Specifically, we employed the complete datasets for AIME2024 and AIME2025,
 831 while for the other datasets we selected 80 samples for evaluation. All model queries were conducted
 832 through APIs, and the temperature parameter (when supported) was set to its default value. To
 833 mitigate the influence of randomness in the evaluation, we reported outcomes that tend toward the
 834 middle of repeated runs. For consistency, we calculate tokens uniformly using the DeepSeek-V3
 835 tokenizer⁶. To enable support for heterogeneous agents, we apply minor modifications to SoM.

836 D ADDITIONAL EXPERIMENTAL RESULTS

837
 838 Table 4: Comparison of reasoning accuracy between our schemes and baselines when $R = 2$.

Benchmarks	GSM-Level4	GSM-Level6	MATH500	Logical Fallacies	AI Crypto	StrategyQA	Avg
FREE-MAD-N	75.00	66.25	65.00	71.25	76.25	68.75	70.42
FREE-MAD-C	75.00	61.25	66.25	71.25	72.50	68.75	69.17
Baseline 1	70.00	58.75	53.75	68.75	58.75	65.00	62.50
Baseline 2	62.50	56.25	58.75	70.00	63.75	65.00	62.71

856 E PROMPT

857
 858 The core prompt for constructing a consensus-free debate is as follows:
 859

860
 861
 862
 863 ⁶https://api-docs.deepseek.com/quick_start/token_usage

864
865**Core Prompt to Mitigating Conformity**866
867

"Since some malicious agents may deliberately disseminate incorrect answers, you must follow the reasoning procedure below and avoid being misled.

868

1. Initial Reasoning

869

Clearly enumerate your initial logical steps and the final conclusion.

870

2. Analysis of Other Agents' Reasoning

871

Specify which agents' reasoning is correct and which contains errors, and identify the exact type of error.

872

You must provide concrete error descriptions rather than general comments. Note that the correct answer may not exist within the current set of answers.

873

3. Comparative Analysis with Your Own Reasoning

874

Examine whether you have made similar mistakes and provide justification for your assessment.

875

4. Final Decision

876

Indicate whether you will revise your conclusion (Yes/No). If yes, explain the reasoning errors; if no, justify why your reasoning stands.

877

5. Additional Requirements

878

You may not rely on the principle of conformity. Majority opinion cannot be used as a basis for adoption.

879

If you cannot definitively determine whether others are correct, you should retain your own conclusion.

880

Each round of reasoning must independently identify errors and must not directly replicate the analysis or conclusions of other agents."

881

882

883

884

885

886

887

888

889

890

891

892

893

894

895

896

897

898

899

900

901

902

903

904

905

906

907

908

909

910

911

912

913

914

915

916

917

918 F RESPONSES
919920 The responses of agents across different debate modes (including conformity and anti-conformity)
921 are compared on various benchmark datasets, under the condition of $R = 1$.
922923 **GSM-Level4**924 **Question:**925 Raymond and Samantha are cousins. Raymond was born 158360 years before Samantha. Raymond had a
926 son at the age of 248911. If Samantha is now 343769 years old, how many years ago was Raymond's son
927 born?
928929 **Conformity:**930 I apologize for the confusion. Upon reviewing the calculations, it seems there is a mistake in the first two
931 solutions. The correct approach is to subtract Samantha's current age from the age of Raymond when he had
932 his son, as the son's birth is what we're looking for in terms of years ago.
933934 Here's the corrected solution:
935936 Raymond's son was born when he was 248911 years old, and Samantha is currently 343769 years old. The
937 difference in their ages is:
938

939
$$343769 - 248911 = 94858 \text{ years}$$

940 This is the number of years that have passed since Raymond had his son. Since we want to find out how
941 many years ago the son was born, we subtract this from Samantha's current age:
942

943
$$343769 - 94858 = 248911$$

944 So, Raymond's son was born 248,911 years ago. The final answer is 248911. This matches the last agent's
945 solution, which is accurate.
946947 **Anti-conformity:**

948 1. Initial Reasoning

949

- Calculate the difference in their ages: Samantha's age (343769) minus the son's birth year (248911).
- Subtract this difference from Samantha's current age to find out how many years ago the son was born.

950 2. Analysis of Other Agents' Reasoning

951

- Agent 1 correctly calculates the difference in ages and then adds the birth difference for Raymond (158360)
952 to find the years until the son's birth. Their error is in assuming that the son's age at the time of Samantha's
953 current age equals zero, which leads to a negative result. This is a sign of miscalculation.
- Agent 2 also performs the same calculation correctly but then subtracts the difference in years to find out
954 how many years ago the son was born, which is correct.
- Agent 3 makes a mistake by subtracting the difference in years directly without considering the birth
955 difference for Raymond. This results in an incorrect answer.

956 3. Comparative Analysis with My Own Reasoning

957

- I agree with both Agent 2 and myself, as we both correctly calculate the years until the son's birth by adding
958 the birth difference for Raymond to the difference in ages.

959

- My reasoning does not contain any errors like the other agents.

960 4. Final Decision

961

- Based on the comparison, my reasoning is correct, and I will not revise my conclusion.

962 5. Additional Requirements

963

- Since both Agent 2 and my own reasoning are correct, there is no need for further revision or additional
964 explanation.

965 Final Answer: 94858 (the same as Agent 2's answer)

966

967

968

969

970

971

972

GSM-Level6

973

974

Question:

975

After scoring 1213526 points, Erin now has 5700191 times more points than Sara, who scored 9. How many points did Erin have before?

976

Conformity:

977

I apologize for the confusion in my previous response. Upon reevaluating the problem, I see that the second agent's solution is correct. The problem states that Sara scored 9 points, and Erin has 5700191 times more points than Sara. So, we can directly calculate Erin's initial score without needing to subtract any points. Here's the corrected reasoning and answer...

978

So, the answer is 51301808. This is the correct number of points Erin had before adding the 5700191 times more points.

979

Anti-conformity:

980

1. Initial Reasoning:

981

I understand that the problem states that Erin currently has 1,213,526 points and is 570,0191 times more points than Sara, who scored 9 points. To find Erin's original score, I need to subtract the additional points she gained from her current score...

982

2. Analysis of Other Agents' Reasoning:

983

- The first agent's solution incorrectly multiplies Sara's score by 570,0191 without considering that this is the additional points, leading to a negative result. This is a mathematical error.

984

- The second agent's solution also makes a similar mistake by directly adding the 570,0191 to Sara's score without adjusting for the additional points.

985

- The third agent's solution appears to be correct, as they subtract the additional points from the current score, which is the proper approach.

986

3. Comparative Analysis with My Own Reasoning:

987

Both the first and second agents' approaches contain the same error: they add the additional points directly to Sara's score instead of subtracting them from Erin's current score. I will follow the correct method from the third agent.

988

4. Final Decision:

989

I will revise my conclusion based on the third agent's reasoning...

990

However, as mentioned earlier, this negative score does not make sense in the context of the problem. It suggests a mistake in the problem setup or the given data. The correct approach should be to subtract the additional points, not add them.

991

5. Additional Requirements: Based on the analysis, I will stick to the corrected method from the third agent and adjust for the error.

992

993

Adjusted for error:

994

995

Erin's original score = $1,213,526 - 5130171$

996

997

Erin's original score = 5130180

998

999

Final Answer: 5130180

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

1021

1022

1023

1024

1025

1026

1027

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

1042

1043

1044

1045

1046

1047

1048

1049

1050

1051

1052

1053

1054

1055

1056

1057

1058

1059

1060

1061

1062

1063

1064

1065

1066

1067

1068

1069

1070

1071

1072

1073

1074

1075

1076

1077

1078

1079

MATH500**Question:**

How many positive whole-number divisors does 196 have?

Conformity:

I apologize for the confusion. The correct answer to the problem is indeed 9, as stated in my initial response.

Here's a clarification for each of the other agent's solutions:

1. The first agent correctly factors 196 into $2^2 \times 7^2$ and then uses the formula for the number of divisors based on the prime factorization. Since there are 2 factors of 2 and 2 factors of 7, the total number of divisors is $(2+1)(2+1) = 3 \times 3 = 9$.

2. The second agent counts the divisors directly by considering the powers of the prime factors. They start with 1 and go up to the square root of 196, which is 14. For each prime factor (2 and 7), they find the possible exponents (0, 1, and 2). For 2, there are 3 divisors (1, 2, and 4), and for 7, there's only 1 (7 itself). Multiplying these gives 5 divisors, but again, they exclude 196 itself, resulting in $5 - 1 = 4$. However, the first agent's solution is correct, and the final answer should be 9.

3. The third agent also finds the correct prime factorization and then counts the divisors. They correctly identify the five divisors mentioned (1, 2, 4, 7, and 14). However, like the first agent, their final answer should be 9.

In summary, all three agents' approaches are valid, but the first agent's solution is the most accurate and consistent with the problem statement, giving the correct answer of 9.

Anti-conformity:

1. Initial Reasoning:

I will follow the standard method to find the number of positive whole-number divisors of 196 by listing all the divisors from 1 to the number itself, including both.

2. Analysis of Other Agents' Reasoning:

- Agent 1 correctly applies the formula for finding the number of divisors based on prime factorization, but they count 196 twice (as 2 and 49), resulting in a total of 9 instead of 8 divisors (excluding 1 and 196).

- Agent 2 makes a mistake by counting 196 twice but stops after 6 divisors, not realizing that 196 is a repeated divisor.

- Agent 3 simply lists the divisors and counts them correctly, identifying 9 divisors.

3. Comparative Analysis with My Own Reasoning:

My reasoning matches Agent 3's, as I also listed the divisors and counted them. I did not repeat any divisor.

4. Final Decision: I will revise my conclusion based on the error analysis. There are 9 divisors, not 8, because 196 is a double-counted divisor. Therefore, the correct answer is 9.

1080

AICrypto

1081

Question:

1083

The following ciphertext was encoded via Caesar cipher with a fixed shift of three letters. Determine the original plaintext: HQFUBSWHG ZRUG.

1084

Options:

1085

- A. ABANDONED WORD
- B. ENCRYPTED TEXT
- C. ABANDONED TEXT
- D. ENCRYPTED WORD
- E. None of the above

1091

Conformity:

1092

Based on the multiple solutions provided by the other agents, all agree that the ciphertext "HQFUBSWHG ZRUG" decodes to "ENCRYPTED WORD" when shifted three positions back in the alphabet using a Caesar cipher. This is a direct result of the encryption process.

1093

Since the question asks for the original plaintext and the decrypted text matches one of the options (D. ENCRYPTED WORD), we can deduce that the original plaintext was encrypted and the correct answer is: D. ENCRYPTED WORD

1094

However, as mentioned in the previous responses, without further context or a list of possible plaintexts, the final answer cannot be definitively stated as "None of the above" because "ENCRYPTED WORD" itself is also a possible option. But given the information provided, the best answer is D. ENCRYPTED WORD.

1095

Anti-conformity:

1096

1. Initial Reasoning:

1097

- The ciphertext "HQFUBSWHG ZRUG" was encrypted using a Caesar cipher with a fixed shift of three letters.
- To find the original plaintext, we need to reverse the encryption by shifting each letter in the ciphertext back by three places.

1098

2. Analysis of Other Agents' Reasoning:

1099

- Agent 1 correctly deciphers the first part as "ENCRYPTED" and the second part as "WORD". However, they claim the original plaintext is "ENCRYPTED WORD" when the correct answer is "ENCRYPTED".
- Agent 2 makes an error in decoding "ZRUG" as "WORD" instead of "ENCRYPTED", and then claims the original is "ABANDONED TEXT", which is incorrect.
- Agent 3 also misinterprets "ZRUG" and concludes the original is "ABANDONED TEXT" without considering the decrypted part.

1100

3. Comparative Analysis with My Own Reasoning:

1101

- I agree with Agent 1 that "HQFUBSWHG" decodes to "ENCRYPTED".

1102

- However, "ZRUG" should also be decrypted to "ENCRYPTED", not "WORD" as all other letters suggest a shift by three places.

1103

4. Final Decision:

1104

- I believe Agent 1's decryption of "HQFUBSWHG" is correct, but their final answer is incorrect due to not considering "ZRUG".

1105

- Since "ZRUG" should also be "ENCRYPTED" after a shift of three, the original plaintext is a combination of "ENCRYPTED" and "ENCRYPTED", which would be "ENCRYPTED ENCRYPTED".

1106

5. Additional Requirements:

1107

- I cannot definitively say that any other agent is completely wrong, but my reasoning is based on the need to consider both parts of the ciphertext together.

1108

Final Answer: **D.ENCRYPTED**

1134
1135**StrategyQA**

1136

Question:

1137

Will the Albany in Georgia reach a hundred thousand occupants before the one in New York?

1138

Conformity:

1139

Based on the solutions provided by the other agents, we have the following options:

1140

1. The first agent's solution assumes that the current populations and projected growth rates are not available and makes a general observation about the relative growth rates and economic factors. This suggests that it's unlikely for Albany, Georgia, to reach 100,000 before Albany, New York, without specific data.

1141

2. The second agent uses October 2021 data, stating that the Albany in Georgia had not yet reached 100,000 and the Albany in New York had around 750,000, so the answer is 'False'.

1142

3. The third agent also relies on hypothetical data and mentions that without specific numbers, the answer cannot be determined.

1143

Considering all the information, it seems that the second agent's solution, based on the October 2021 data, provides a concrete basis for answering the question. At that time, the Albany in Georgia did not have a population of 100,000, and the Albany in New York did. Therefore, given this information, the answer would be:

1144

True - The Albany in Georgia would not have reached a hundred thousand occupants before the one in New York, as per the data from October 2021.

1145

Anti-conformity:

1146

1. Initial Reasoning:

1147

- The question asks if the Albany in Georgia will reach 100,000 occupants before the one in New York.

1148

- The given solutions involve analyzing current population data, considering growth rates, and comparing the two cities' populations.

1149

- The first agent's solution correctly states that with the current difference in population, it's unlikely for Albany, Georgia, to reach 100,000 before Albany, New York. They provide numerical estimates and use the word "unlikely," indicating a high probability of this not happening.

1150

- The second agent's solution assumes constant growth rates and uses simple arithmetic to project future populations. They also conclude that it's unlikely for Albany, Georgia, to reach 100,000 before New York, based on the provided growth rates.

1151

- The third agent's solution acknowledges the lack of specific numbers but suggests an "unknown" answer because of the missing data. They imply that without exact figures, a definitive answer cannot be given.

1152

2. Analysis of Other Agents' Reasoning:

1153

- All three agents agree that the current population gap and steady growth rates point to a likelihood of Albany, Georgia, not reaching 100,000 before Albany, New York. Their error, if any, lies in the assumption of constant growth rates, which might not necessarily hold in real life due to various factors like economic cycles, natural disasters, or policy changes.

1154

- The first and second agents provide numerical estimates, which are based on available data and logical reasoning. They do not commit any significant logical errors.

1155

3. Comparative Analysis with My Own Reasoning:

1156

My reasoning aligns with the majority of the agents, as I too conclude that it's unlikely for Albany, Georgia, to reach 100,000 before the one in New York due to the existing population disparity and steady growth patterns. I do not find any significant errors in their reasoning.

1157

4. Final Decision:

1158

Yes, I will revise my conclusion based on the collective reasoning of the other agents. The main error in their analysis is the assumption of constant growth rates, which might not accurately represent the complexities of real-world population dynamics. However, since none of them explicitly acknowledge this, and my own reasoning supports their overall conclusion, I will maintain the same answer: **False**.

1188 **G STATEMENTS**
11891190 The methodology presented in this work is intended exclusively for scientific research. In this paper,
1191 LLMs are used to polish the writing.
11921193 **H ANALYSIS AND DISCUSSION**
11941195 **H.1 COMPATIBILITY**
11961197 FREE-MAD is highly compatible with existing MAD frameworks. In the debate stage, FREE-MAD
1198 supports arbitrary debate structures, including Sparse MAD configurations where interactions are
1199 preserved only among a subset of agents. In the decision stage, our score-based decision mechanism
1200 is completed during the execution of the debate itself and does not interfere with any additional
1201 decision protocols applied afterward. As a result, it can coexist with majority voting, LLM-as-a-Judge,
1202 and other decision-making strategies.
12031204 **H.2 LIMITATIONS**
12051206 In this paper, we construct a general MAD framework that incorporates a controllable conformity
1207 debate process together with a score-based mechanism that determines the final decision across all
1208 debate outcomes. Due to budget constraints, in the weighting coefficients \mathcal{W} of the score mechanism,
1209 we adopt a single set of coefficients derived from the theoretical analysis of debate logic, which
1210 we argue to be realistic in actual debate scenarios. While this setting already leads to a significant
1211 improvement in the reasoning accuracy of MAD, we believe that alternative configurations of the
1212 weighting coefficients may further enhance both the accuracy and robustness of FREE-MAD.
12131214 In addition, several related works have compared MAD against single-agent reasoning approaches
1215 to demonstrate that MAD outperforms traditional single-agent methods. Since a substantial body
1216 of research Yang et al. (2025) has already established this conclusion, and given that the baseline
1217 method adopted in this paper has also been shown to outperform single-agent approaches, we omit
1218 repeating such results here for brevity, although we have performed the corresponding verification.
12191220 **H.3 FUTURE WORK**
12211222 Building on the discussion of limitations in Section H.2, we plan to further investigate the impact of
1223 different weighting configurations \mathcal{W} on the score-based decision stage, with the goal of identifying
1224 coefficient settings that can support stronger reasoning accuracy and robustness of MAD. In addition,
1225 we will construct more heterogeneous MAD systems by incorporating a broader range of LLMs and
1226 more challenging benchmarks, thereby further validating the generality of the proposed framework.
1227 For example, we intend to examine the performance of MAD instantiated with reasoning LLMs such
1228 as DeepSeek-R1 (Guo et al., 2025) on the HLE benchmark (Phan et al., 2025). Regarding the study
1229 of MAD’s security, we will employ a wider variety of attacks, such as prompt injection attacks (Liu
1230 et al., 2024b), to provide a more comprehensive evaluation of the framework.
12311232
1233
1234
1235
1236
1237
1238
1239
1240
1241